**NLU Assignment 2 Report**

Mert Akkor

**Q1:** For the first part of this assignment, it was asked to evaluate token-level and chunk-level metrics for Spacy NER on the CoNLL 2003 dataset. To start with, corpus is given as an input to Spacy of an array of words and make it into tokenization format. However, format of the named entity labels that are given by Spacy did not match with the ones in the dataset of CoNLL. To solve this, I created a dictionary that maps Spacy NER to the ones in the dataset. After mapping, I evaluated the token-level by using ‘classification\_report’ from scikitlearn, and for the chunk-level evaluation I used the evaluate function that is already provided to us in conll.py.

**Q2:** In the second part it was asked to group recognized name entities by using noun\_chunk. To do this, it is possible to identify which entity belongs to which chunk without losing the entities that does not belong to any chunk, by iterating over the named entities in the parsed document. It compares the current entity in the parsed document and the entity in the noun\_chunk and entity is added to the chunk if they are equal. Otherwise, entity is added to new created chunk. (I figured out I did it wrong so deleted from the code)

**Q3:** As the possible post-processing step, third exercise ask to fix segmentation steps.